WEB APPENDIX

Some analytic remarks on the relationship between the variance of the IV estimate and the size of the subset on which the exposure is measured

An analytic formula relating the power of a Mendelian randomization analysis to parameter values representing sample size, subset sample size, strength of the genetic IV, causal effect of exposure on outcome, and degree and direction of confounding of the exposure-outcome association is unlikely to be succinct whether in the complete-data, subsample or two-sample situation. In this appendix, we investigate the variance of the Wald estimate in complete-data, and then subsample and two-sample settings, giving analytic insight which could form the basis of an informed power calculation for an investigator designing a specific Mendelian randomization experiment.

The Wald IV estimator ($\hat{\beta}_{IV}$) is the ratio of the reduced form estimate (the coefficient in the regression of the outcome on the IV, $\hat{\beta}_{GY}$) to the first-stage estimate (the coefficient in the regression of the exposure on the IV, $\hat{\beta}_{GX}$). A further expression of the variance of the Wald estimator can be calculated using the delta method (1):

$$var(\hat{\beta}_{IV}) = var\left(\frac{\hat{\beta}_{GY}}{\hat{\beta}_{GX}}\right) \approx \frac{var(\hat{\beta}_{GY})}{\hat{\beta}_{GX}^{2}} + var(\hat{\beta}_{GX})\frac{\hat{\beta}_{GY}^{2}}{\hat{\beta}_{GX}^{4}} - 2\,cov(\hat{\beta}_{GY}, \hat{\beta}_{GX})\frac{\hat{\beta}_{GY}}{\hat{\beta}_{GX}^{3}}$$

We use this expression here as it divides the variance into three convenient terms which can be discussed individually, it is asymptotically equivalent to the two-stage least squares variance in large samples, and its behavior should be similar to that of other variance estimates. The expression is rarely used in practice, as it requires knowledge of the covariance of the estimates $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$, which is related to the quantity which is being estimated, and as it makes

the assumption that the Wald estimator is approximately normally distributed, which is not true in small samples. In the absence of confounding, if the reduced form and first-stage estimates are taken on the same population, the correlation between these parameters is approximately equal to the observational correlation between the exposure and outcome.

Without loss of generality, we scale the IV (*G*), exposure (*X*) and outcome (*Y*) so that each of them has unit variance. We initially assume that the association between *X* and *Y* is unconfounded, and that data are available on *G*, *X*, and *Y* in a sample of *n* individuals.

We have:

$$\hat{\beta}_{GX} = \sqrt{R^2_{GX}}, \hat{\beta}_{GY} = \sqrt{R^2_{GY}} \approx \sqrt{R^2_{GX}} \times \sqrt{R^2_{XY}}$$

$$var(\hat{\beta}_{GX}) = var(\hat{\beta}_{GY}) = \frac{1}{n}$$

$$cov(\hat{\beta}_{GY}, \hat{\beta}_{GX}) = \frac{1}{n} \times \sqrt{R^2_{XY}}$$

where R^2_{GX} is the coefficient of determination is the regression of *X* on *G*. We can approximate the variance of the Wald estimator as:

$$\frac{1}{nR^2_{GX}} + \frac{R^2_{GX} \times R^2_{XY}}{nR^4_{GX}} - 2\frac{\sqrt{R^2_{XY}} \times \sqrt{R^2_{GX}} \times \sqrt{R^2_{XY}}}{n\sqrt{R^2_{GX}}^3}$$

$$= \frac{1}{nR^2_{GX}}(1 + R^2_{XY} - 2R^2_{XY})$$

We have chosen not to simplify the expression further to maintain the three terms in the equation to comparative purposes. We note that the coefficient of determination R^2_{XY} is typically of the order of 0.05-0.2 for most exposures used in Mendelian randomization. This means that the first term in this expression for the variance is typically 5 to 20 times the size of the other terms. If the *X*-*Y* association is thought to be partially driven by confounding, then the true correlation of the

estimates $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$ is closer to the null than the observational correlation between *X* and *Y*, and the third term in the expression further attenuates.

In a subsample analysis, we assume that data on *G* and *Y* are available on n_Y individuals and data on *X* are available on a subset n_X ($\leq n_Y$) individuals. The expression of the variance of the Wald estimator is:

$$\frac{1}{n_Y R_{GX}^2} + \frac{R_{GX}^2 \times R_{XY}^2}{n_X R_{GX}^4} - 2\frac{cor\left(\hat{\beta}_{GY}, \hat{\beta}_{GX}\right) \times \sqrt{R_{GX}^2} \times \sqrt{R_{XY}^2}}{\sqrt{n_X n_Y} \times \sqrt{R_{GX}^2}^3}$$

$$= \frac{1}{R_{GX}^2}\left(\frac{1}{n_Y} + \frac{R_{XY}^2}{n_X} - 2\frac{cor\left(\hat{\beta}_{GY}, \hat{\beta}_{GX}\right) \times \sqrt{R_{XY}^2}}{\sqrt{n_X n_Y}}\right)$$

As the estimates of $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$ are no longer derived from the same population, the correlation between the parameters will reduce. Ignoring this term (which would be precisely zero in a two-sample analysis), setting $R_{XY}^2 = 0.1$ and $n_X = n_Y$, the first term is 10 times the magnitude of the second term. This means that decreasing the subset size n_X to 10% of n_Y would only approximately double the variance of the Wald estimator.

The relationship between the variance of an estimator and power is not linear, and a doubling of the variance may or may not have an appreciable impact on power. However, we have demonstrated, subject to several approximating and simplifying assumptions, that substantial reduction of the subsample on which the exposure is measured does not necessarily increase the variance of the IV estimator by the same proportion.

Fieller's theorem

If the regression coefficients in the ratio method $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$ are assumed to be normally distributed, critical values and confidence intervals for the estimator may be calculated using

Fieller's theorem (2). For this, we need the correlation between $\hat{\beta}_{GY}$ and $\hat{\beta}_{GX}$, which is generally assumed to be zero (3). If the standard errors are $se(\hat{\beta}_{GY})$ and $se(\hat{\beta}_{GX})$ and the sample size is *N*, then we define:

$$f_0 = \hat{\beta}_{GY}{}^2 - t_N(0.975)^2\, se(\hat{\beta}_{GY})^2$$

$$f_1 = \hat{\beta}_{GX}{}^2 - t_N(0.975)^2\, se(\hat{\beta}_{GX})^2$$

$$f_2 = \hat{\beta}_{GX}\, \hat{\beta}_{GY}$$

$$D = f_2{}^2 - f_0 f_1$$

where $t_N(0.975)$ is the 97.5th percentile point for a *t*-distribution with *N* degrees of freedom.

If $D > 0$ and $f_1 > 0$, then the 95% confidence interval is the interval from $\frac{f_2-\sqrt{D}}{f_1}$ to $\frac{f_2+\sqrt{D}}{f_1}$. The confidence interval is more likely to be a closed interval like this if we have a "strong" instrument, that is an instrument which explains a large proportion of the variation of the exposure in the population.

If $D < 0$, then there is no interval which covers the true parameter with 95% confidence. This occurs when there is little differentiation between both the exposure and outcome distributions in the genetic subgroups, and so an estimate corresponding to any size causal association is plausible.

If $D > 0$ and $f_1 < 0$, then the 95% confidence interval runs is the union of two intervals from minus infinity to $\frac{f_2+\sqrt{D}}{f_1}$and then from $\frac{f_2-\sqrt{D}}{f_1}$to plus infinity. All possible values are included in the interval except those between $\frac{f_2+\sqrt{D}}{f_1}$and $\frac{f_2-\sqrt{D}}{f_1}$. The interpretation is that the IV estimate is compatible with infinity, but not compatible with a finite association of a given magnitude.

To summarize, Fieller's theorem gives confidence intervals that have one of three possible forms:

i. The interval may be a closed interval $[a, b]$,

ii. The interval may be the complement of a closed interval $(-\infty, b] \cup [a, \infty)$,

iii. The interval may be unbounded.

where $a = \frac{f_2 - \sqrt{D}}{f_1}$, and $b = \frac{f_2 + \sqrt{D}}{f_1}$.

Stata code for conducting subsample (or two-sample IV) analysis using seemingly unrelated regression (SUR) and the delta method

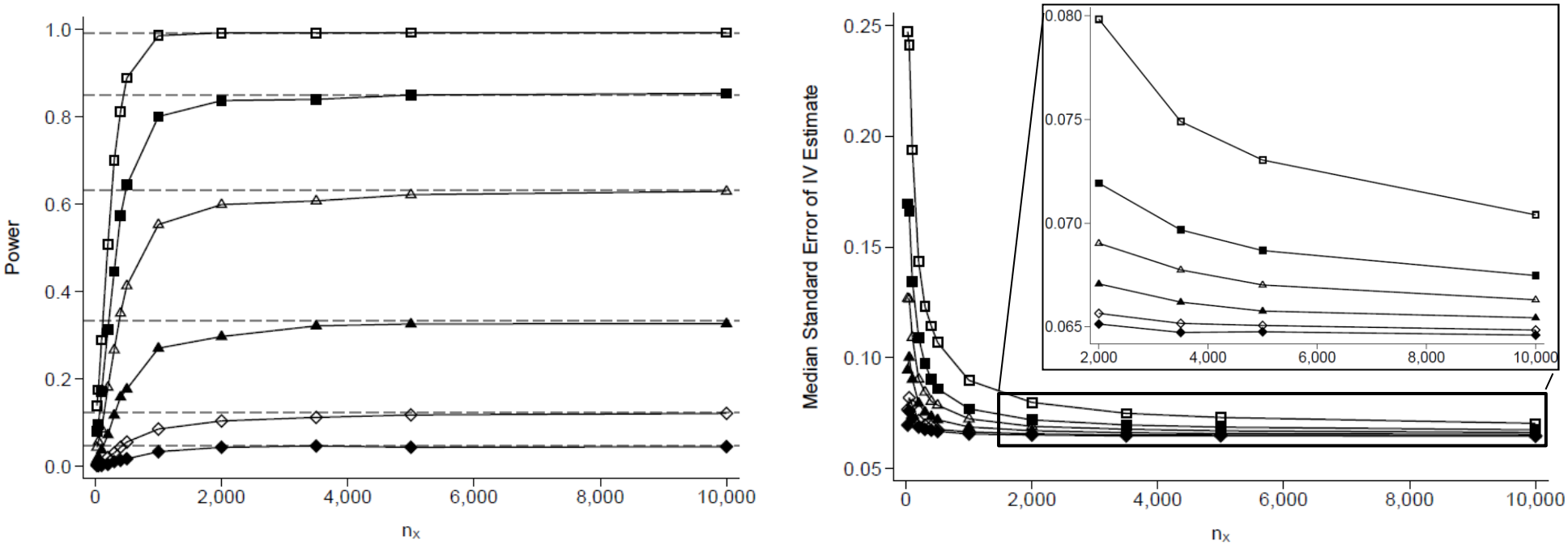
```
/***** g is the instrument, x is the risk factor, y is the outcome *****/
/****participants with data on x are observations 1 through nX ****/

/*fit the reduced form regression model using all observations and store results*/
regress y g
est store GY
/*fit the first-stage regression model using nX observations and store results*/
regress x g in 1/nX
est store GX
/* combine results from different estimation commands and obtain MR/IV estimate*/
suest GY GX
nlcom ([GY_mean]g)/([GX_mean]g) , post
```

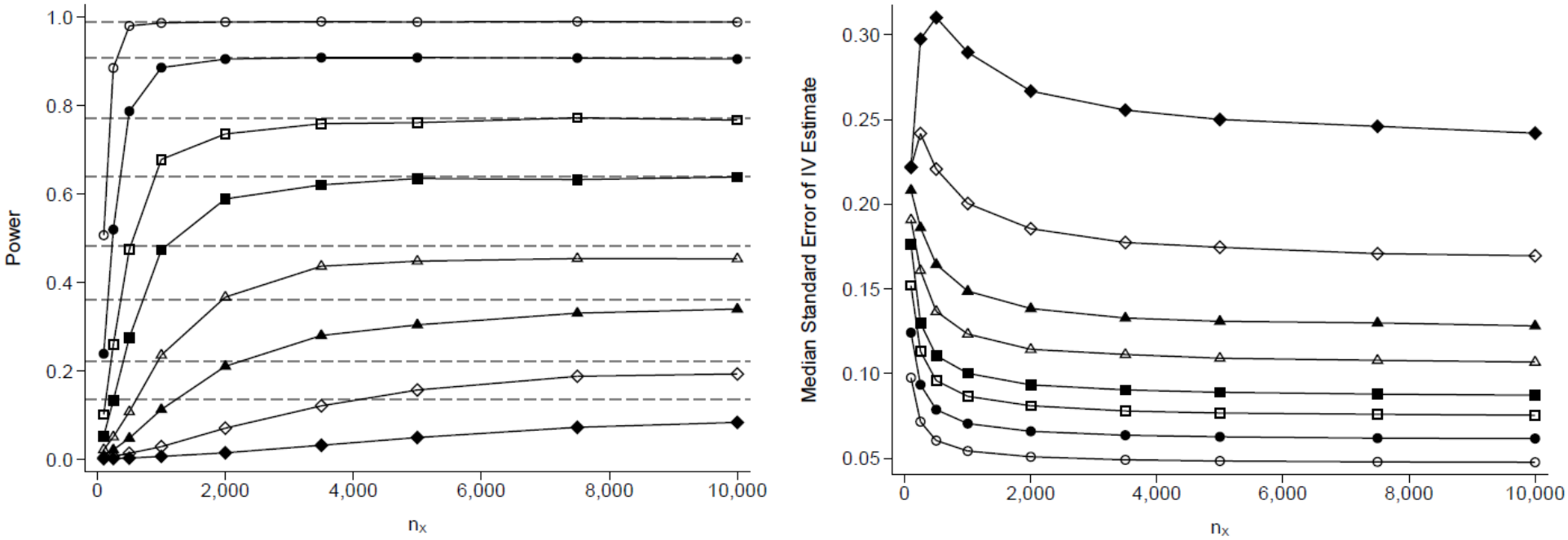
REFERENCES

1. Thomas DC, Lawlor DA, Thompson JR. Re: "Estimation of bias in nongenetic observational studies using 'Mendelian triangulation' by Bautista et al." *Ann Epidemiol* 2007;17:511–513.
2. Fieller E. Some problems in interval estimation. *J Roy Stat Soc Ser B (Stat Methodol)* 1954;16:175–185.
3. Minelli C, Thompson JR, Tobin MD, et al. An integrated approach to the meta-analysis of genetic association studies using Mendelian randomization. *Am J Epidemiol* 2004;160:445–452.

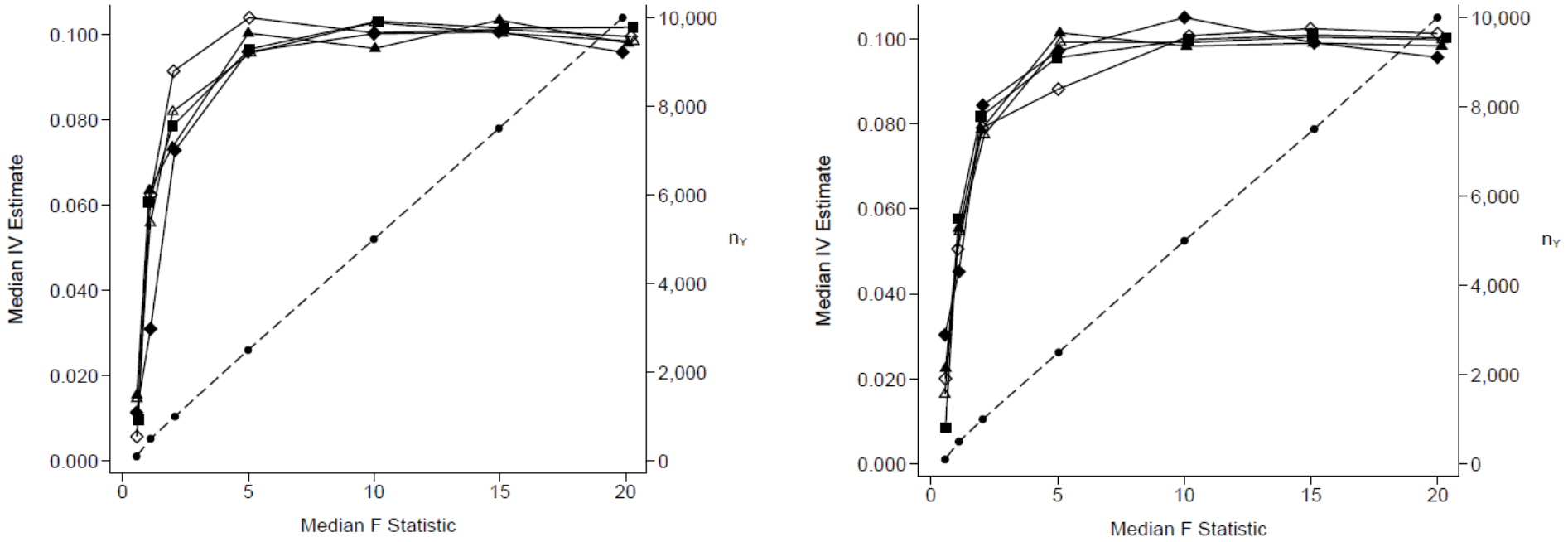
Web Figure 1. Power (left) and median standard errors (right) for the two-sample IV estimate for different values of the causal effect size (β_{XY}) and the sample size of the first-stage regression (n_X), with a strong IV (R^2 = 0.025), a sample size for the reduced form regression (n_Y) of 10,000, and a confounding variable with equal effects on *X* and *Y* ($\beta_{UX} = \beta_{UY}$ = 0.2). β_{XY} values are 0.0 (filled diamond), 0.05 (open diamond), 0.1 (filled triangle), 0.15 (open triangle), 0.2 (filled square), and 0.3 (open square).



Web Figure 2. Power (left) and median standard errors (right) for the two-sample IV estimate for different values of the first-stage R^2 and the sample size of the first-stage regression (n_x) with a constant causal effect size ($\beta_{XY} = 0.2$), a sample size for the reduced form regression (n_Y) of 10,000, and a confounding variable with equal effects on *X* and *Y* ($\beta_{UX} = \beta_{UY} = 0.2$). First-stage R^2 values are 0.002 (filled diamond), 0.004 (open diamond), 0.007 (filled triangle), 0.01 (open triangle), 0.0015 (filled square), 0.2 (open square), 0.03 (filled circle), 0.05 (open circle).



Web Figure 3. Bias in the two-sample IV estimate for confounded (left) and unconfounded (right) scenarios for different values of the average first-stage *F* statistic and the relative size of the sample used in the first-stage regression (n_X:n_Y) with a constant causal effect size ($\beta_{XY} = 0.1$). Values for n_X:n_Y are 1 (filled diamond), 0.75 (open diamond), 0.5 (filled triangle), 0.25 (open triangle), and 0.1 (filled square). The sample size for the reduced form equation (n_Y, on the right vertical axis) is shown as dots connected with a dotted line.



Web Table 1. Power estimates from simulation 1 the delta method under the null hypothesis of no effect ($\beta_{XY} = 0.0$)

Subsample IV					Two-sample IV				
n_X	Median Beta	Median SE	Median *F*	Power	n_X	Median Beta	Median SE	Median *F*	Power
30000	0.000	0.064	769.9	0.0476	30000	0.000	0.065	768.6	0.048
20000	0.000	0.065	512.6	0.0512	20000	-0.001	0.065	512.4	0.053
10000	0.000	0.065	256.3	0.0461	10000	0.001	0.065	256.4	0.045
5000	0.000	0.065	128.3	0.0498	5000	0.000	0.065	128.3	0.044
3500	0.001	0.065	89.8	0.0479	3500	-0.001	0.065	90.0	0.047
2000	0.001	0.065	51.1	0.0397	2000	0.001	0.065	51.3	0.044
1000	0.000	0.065	25.9	0.032	1000	0.000	0.066	25.6	0.033
500	0.000	0.066	12.8	0.0193	500	0.000	0.067	12.7	0.017
400	-0.001	0.067	10.3	0.0149	400	0.000	0.067	10.2	0.014
300	-0.001	0.068	7.7	0.0096	300	-0.001	0.068	7.6	0.011
200	-0.001	0.069	5.1	0.0039	200	0.001	0.069	5.1	0.004
100	-0.001	0.073	2.5	0.0019	100	-0.001	0.073	2.5	0.002
50	0.001	0.075	1.3	0.002	50	0.000	0.076	1.3	0.002
25	0.001	0.070	0.8	0.001	25	-0.001	0.070	0.8	0.002

Simulated data sets consist of 10,000 individuals with data on *G* and *Y* and n_X individuals with data on *G* and *X*. A confounding variable *U* has effect of 0.2 on both *X* and *Y*. The first-stage R^2 is 0.025.